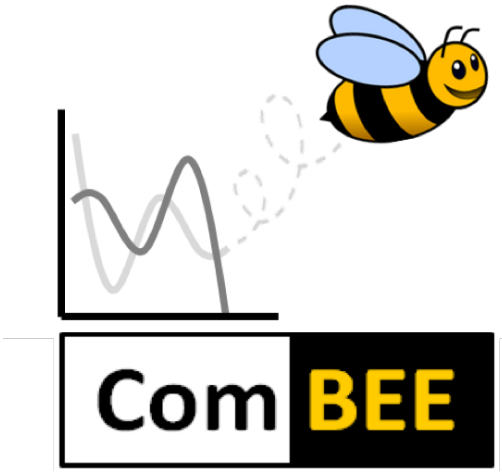
“Best” Practices in Metagenomic Binning

Elizabeth McDaniel

PhD Student, Microbiology Doctoral Training Program

McMahon Lab, Departments of Bacteriology & Civil and Environmental Eng.

@lizilla93
@mcmahonlab
@ComBEE_UW



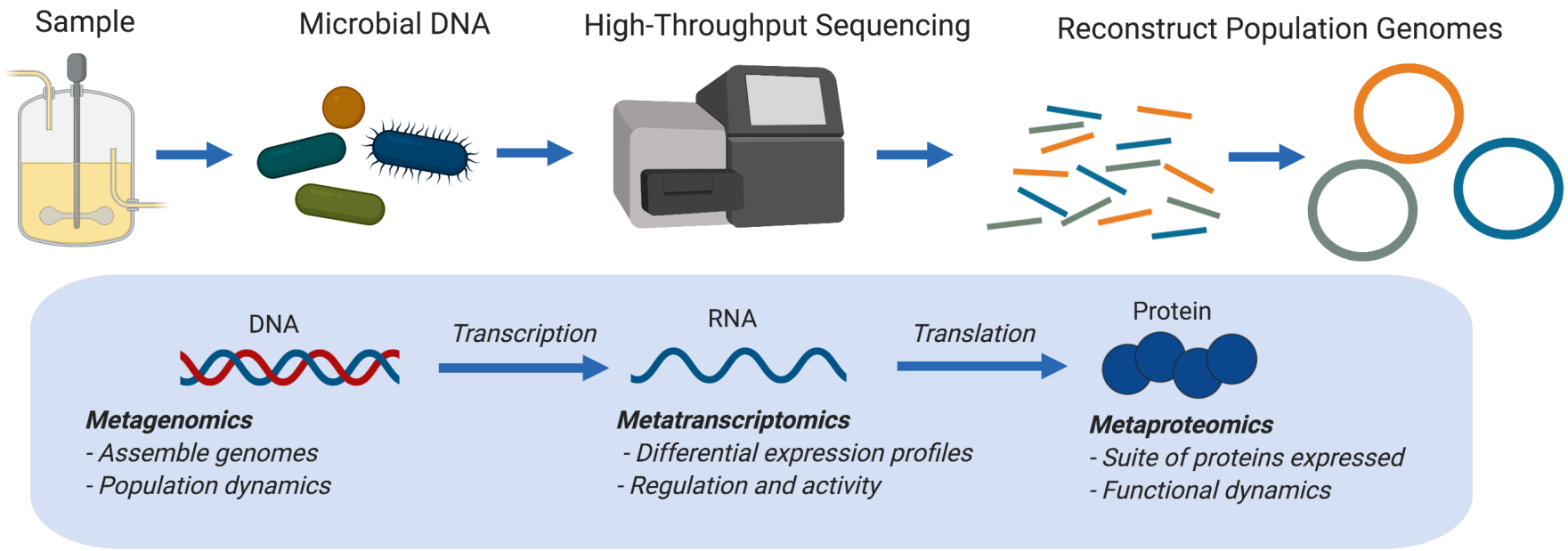
Announcements

- Dr. Tammy Steeves Seminar: *Bicultural approaches for enhancing resilience in Aotaearoa New Zealand's bioheritage*

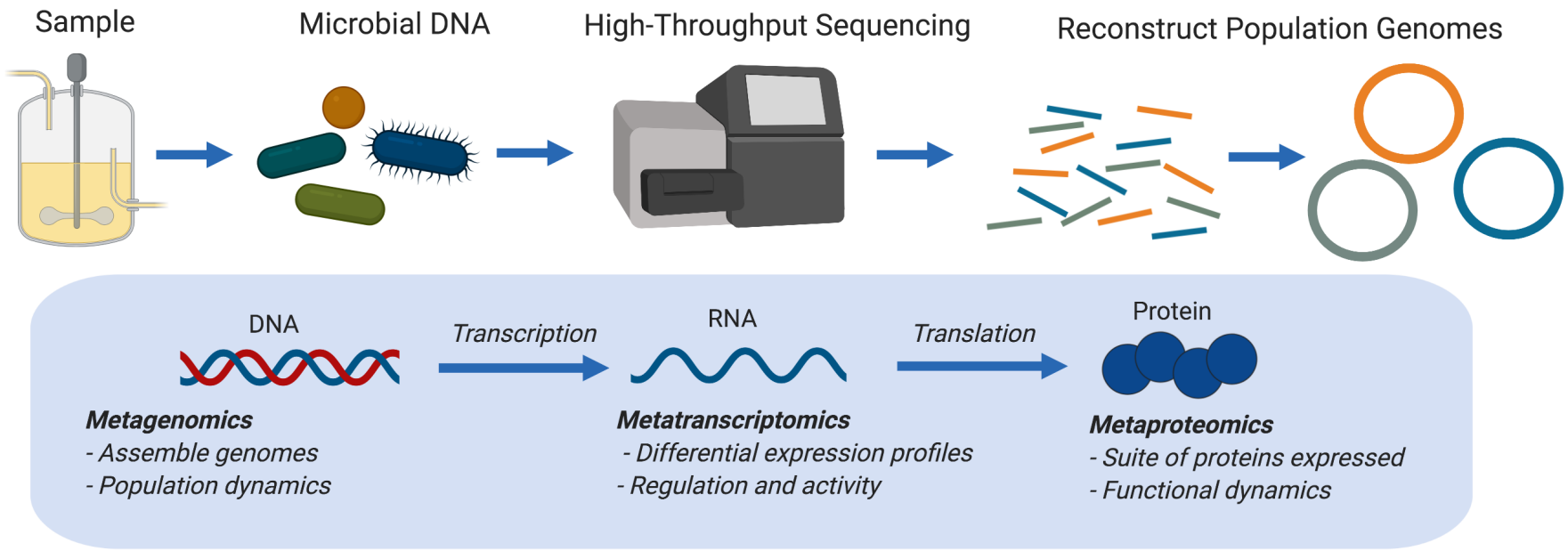
- Last 'Omics Study Group of the semester next week: *Annotation Methods and Databases*

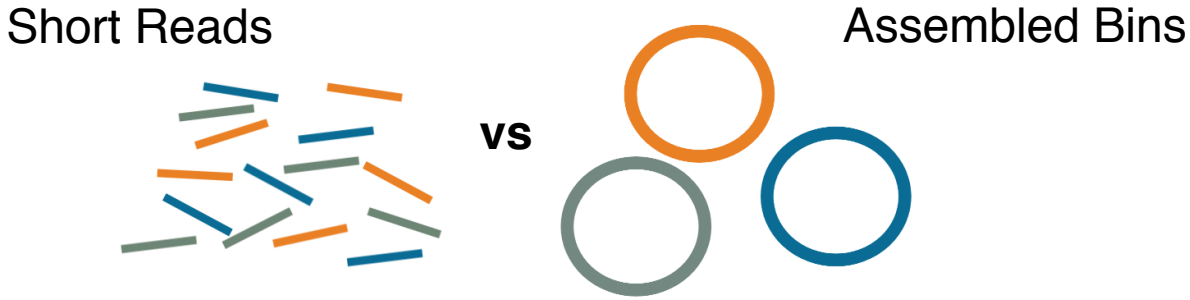
- Actively looking for study group/session leaders and topic ideas for next semester

What is Metagenomics?



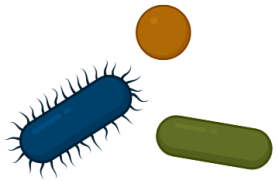
What is Metagenomics?



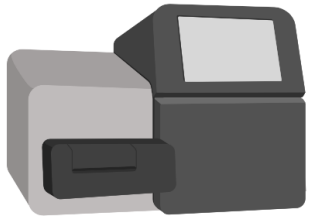


What is Binning?

Microbial DNA



High-Throughput Sequencing



Short Reads

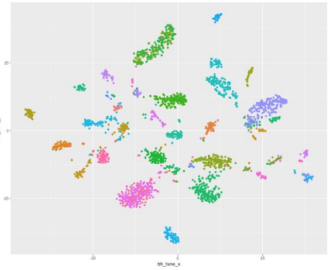
Assemble Contigs

Clustering

Sequence Composition

kmer frequency (tetranucleotide frequencies)

Coverage Composition

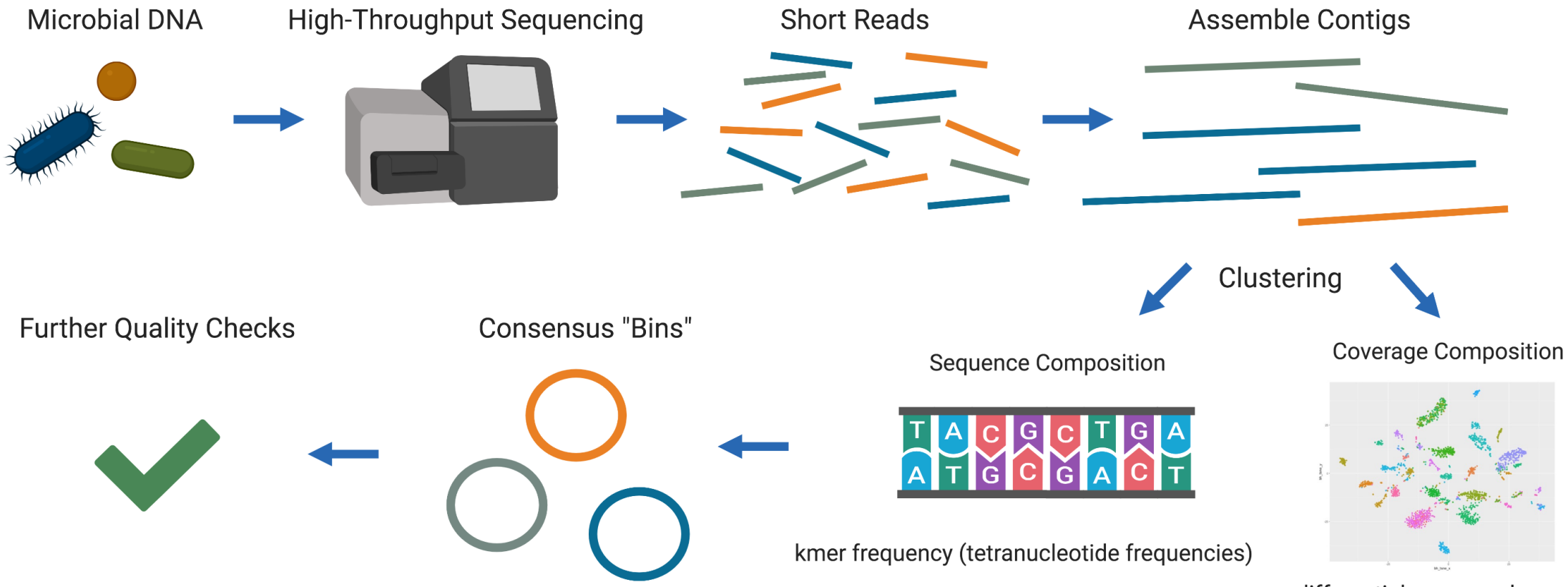


differential coverage based clustering of contigs

Consensus "Bins"

Further Quality Checks

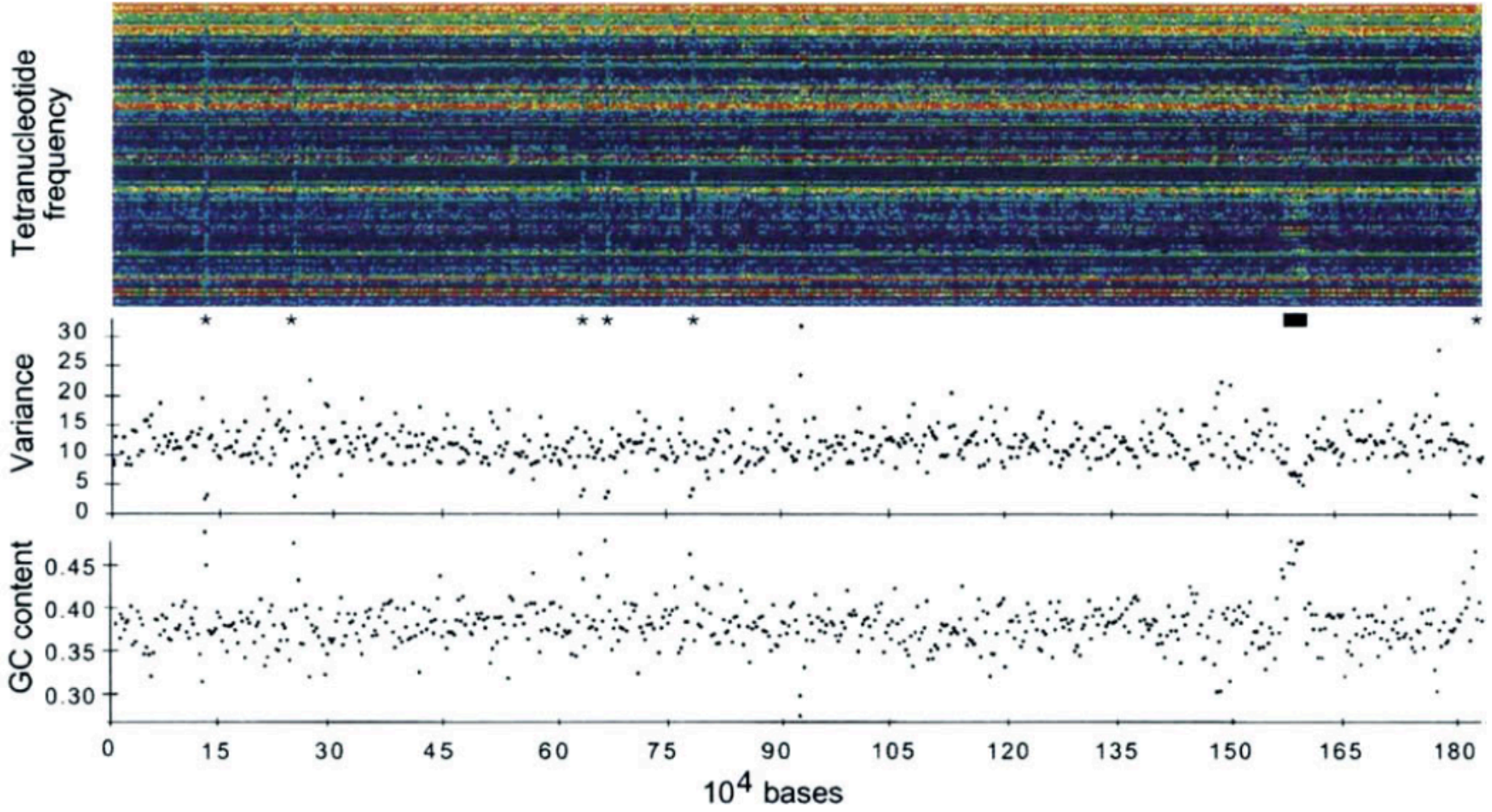
What is Binning?



Other names for bins:
MAGs: Metagenome Assembled Genomes
Population Genomes (depending on mapping % identity for species/strain resolved bins)
Consensus Genomes
GFM: Genomes from Metagenomes

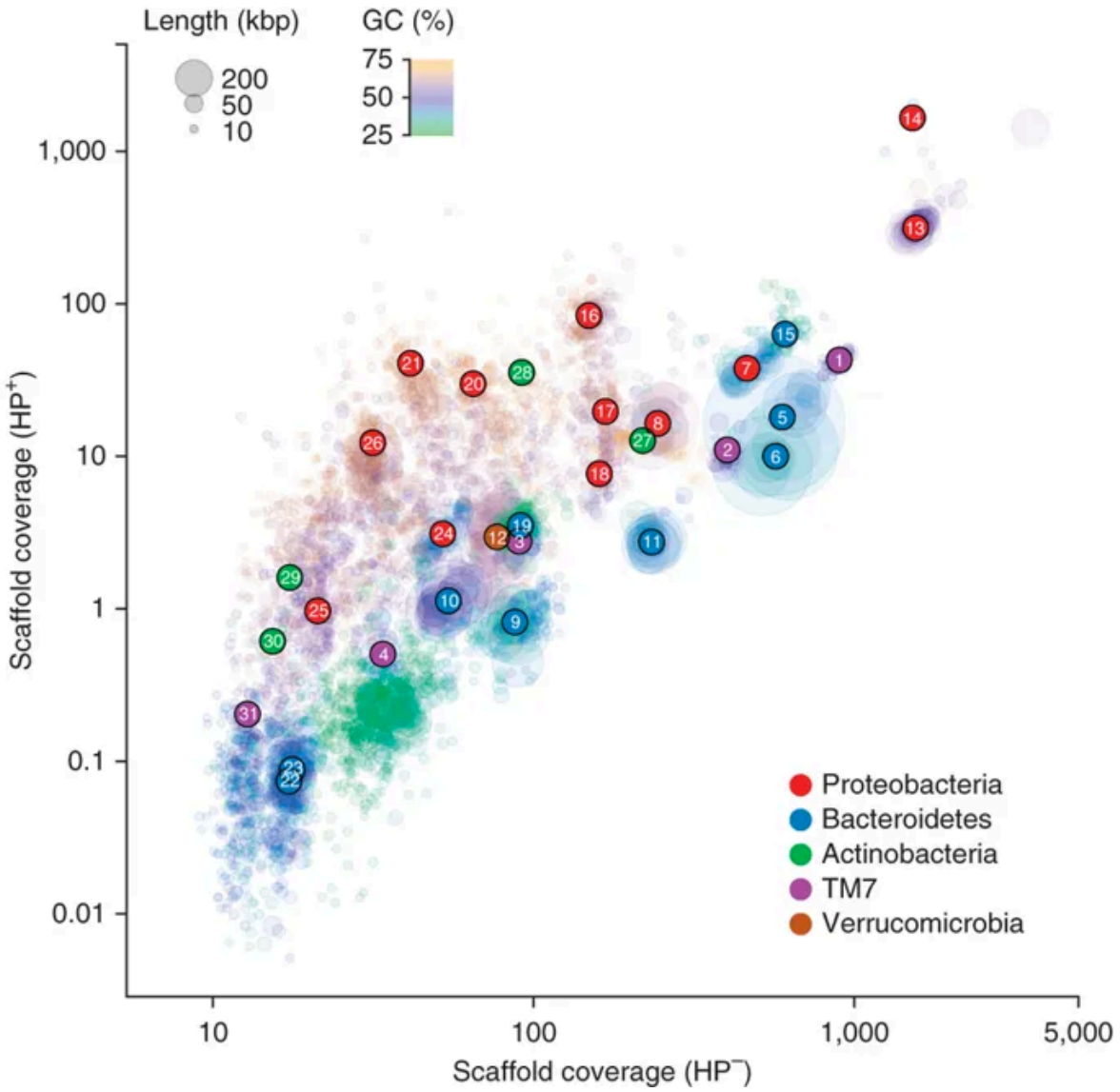
What is Composition-Based Binning?

Non-random sequence signatures present throughout the entire genome, non-linearly related to GC content



Noble et al. 1998

What is Coverage Based Binning?

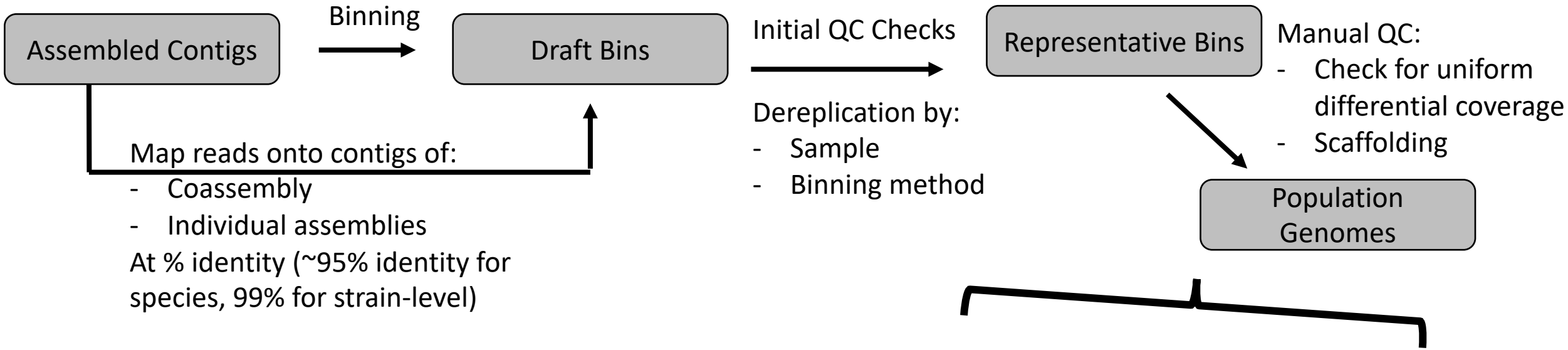


Temporal or spatial sampling of a microbial community from a given environment reveals underlying differences in abundance or differential coverage of populations

Differences in extraction protocols (hot vs cold phenol extraction) = differences in abundance for differential coverage based clustering of contigs

Albertsen et al. 2013

What does a typical binning workflow entail?



Detailed Classification

- Confirm classification at multiple levels
- Make trees compared to reference genomes

Functional Predictions

- Metabolic reconstructions
- Specific functions: secondary metabolites, CAZymes

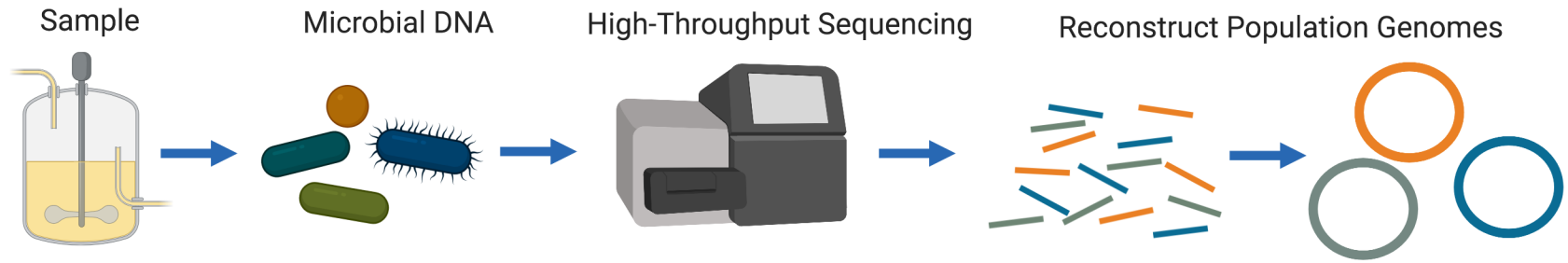
Population Comparisons

- Pangenomics of core vs accessory gene content
- Dynamics specific pops over time/space

Genome Resolved 'Omics

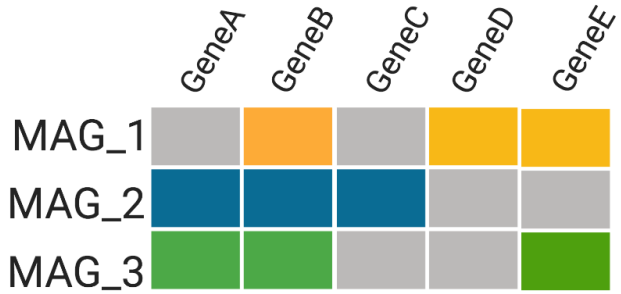
- Metatranscriptomics for more specific activity, gene regulation
- Metaproteomics functional activities, interactions

Why should I spend my time doing this?

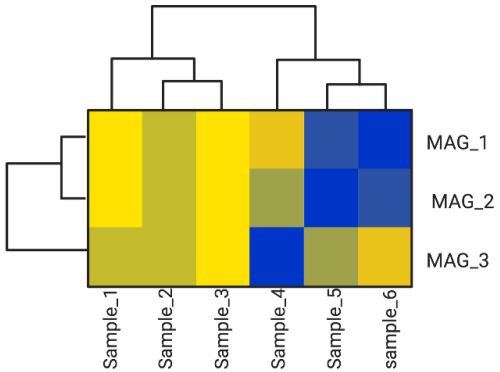


Population Genomes

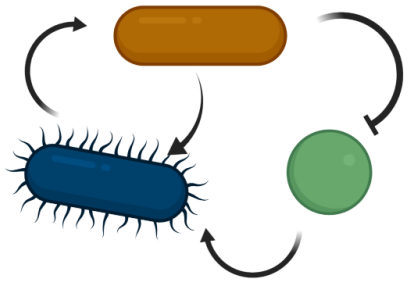
Predicted Functions



Expressional Activity



Putative Interactions



(Mostly) Automatic Binning Algorithms

- MetaBAT & MetaBAT2
- CONCOCT
- Maxbin & Maxbin2
- ABAWACA
- COCACOLA
- BinSanity
- Autometa
- ESOMS (emergent self-organizing maps)
- mmgenome2
- Anvi'o Binning Workflow

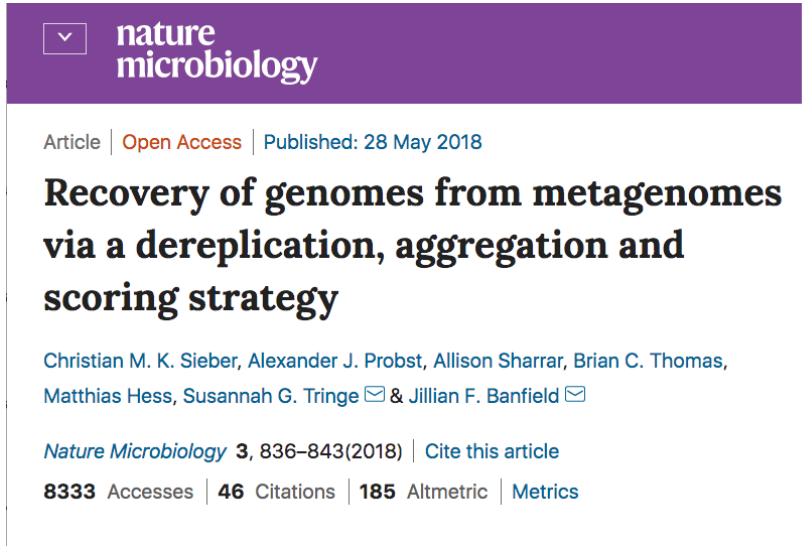
….. plus many more

Dereplication of Samples & Binning Methods

Dereplication based on two parameters: assembly method and binning method

If you choose to assemble all your samples individually instead of a giant coassembly, can reciprocally map all metagenomic reads to all samples, and dereplicate based on the “best” bin cluster with ***drep***

There are also differences in binning algorithms, and for mysterious reasons sometimes certain binning programs work best for specific environments, so becoming a best practice to try multiple binning approaches and dereplicate among those to get the “best” representative bin cluster, also with,. originally in ***DAStool*** but also implemented in ***drep***



Quality Checks

Assessing the overall quality of your MAGs, usually by estimates of completion and redundancy (contamination).

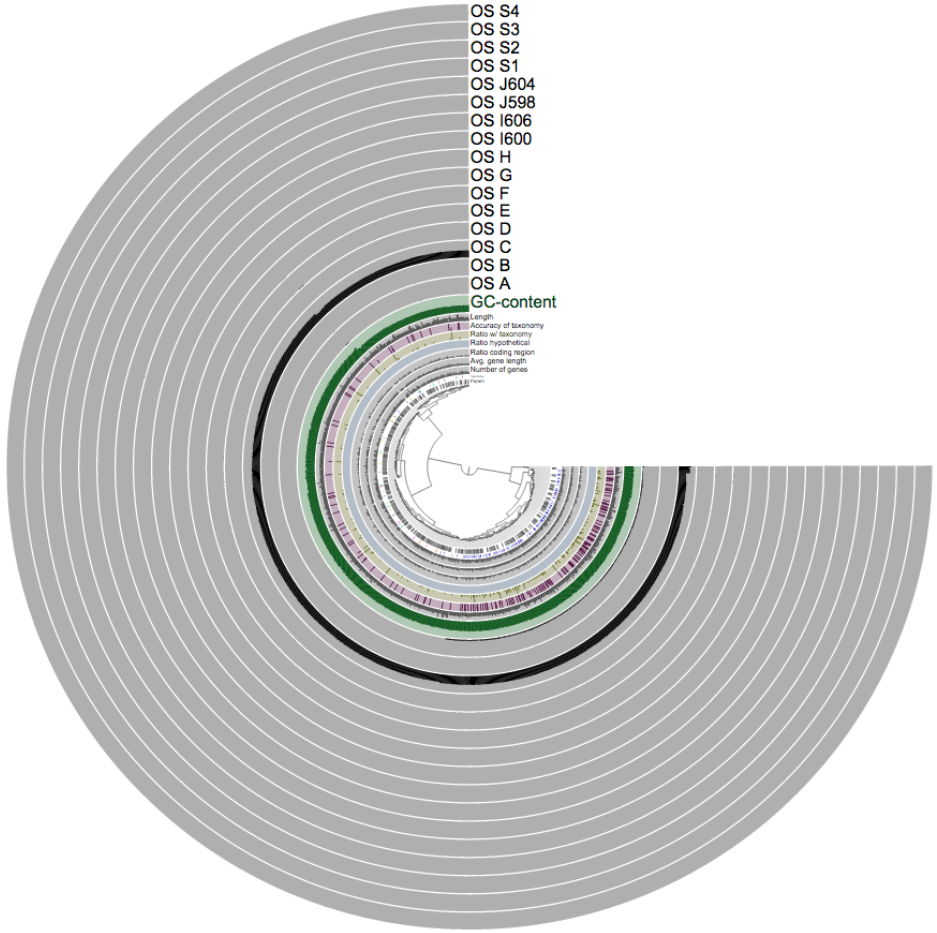
CheckM = ”universal” single-copy core genes, should theoretically be in all bacteria/archaea once

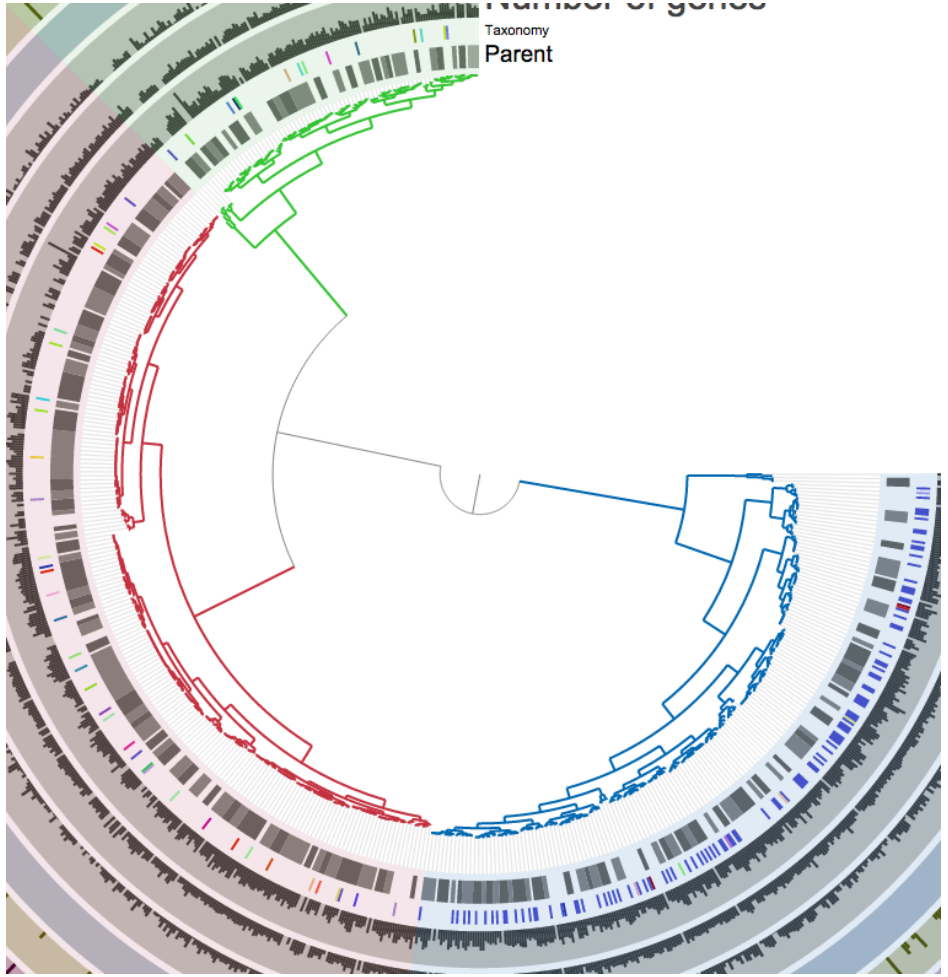
Australian Centre for Ecogenomics

$$\text{Completeness} = \frac{\text{\# of single copy genes found}}{\text{\# of single copy genes searched for}}$$

$$\text{Redundancy} = \frac{\text{\# of single copy genes found more than once}}{\text{\# of single copy genes searched for}}$$

Quality Checks





Quality Checks

Medium Quality MAG:
> 50% complete
< 10% redundant

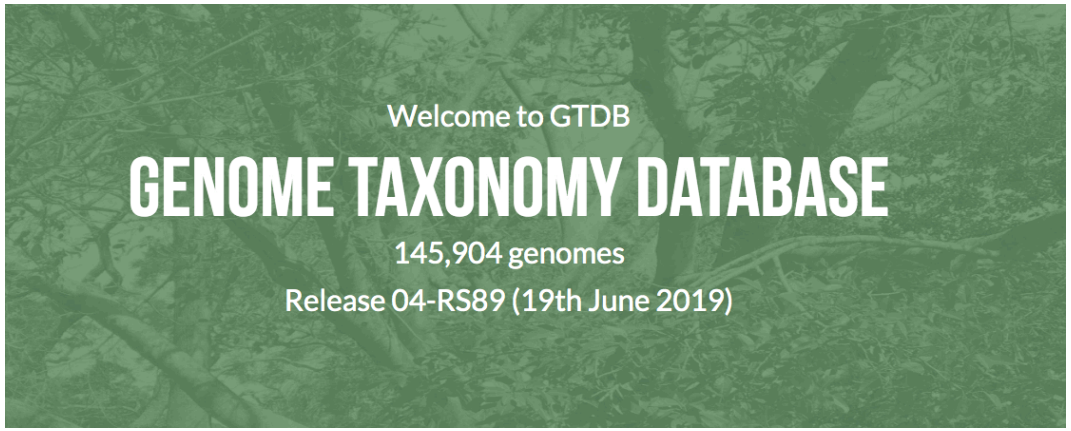
High Quality MAG:
> 95% complete
< 5% redundant
Presence of 16S, 23S, and 5S ribosomal subunits
> 18 tRNAs
(check the latter with Barrnap, Infernal)

Perspective | Open Access | Published: 01 August 2017

Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea

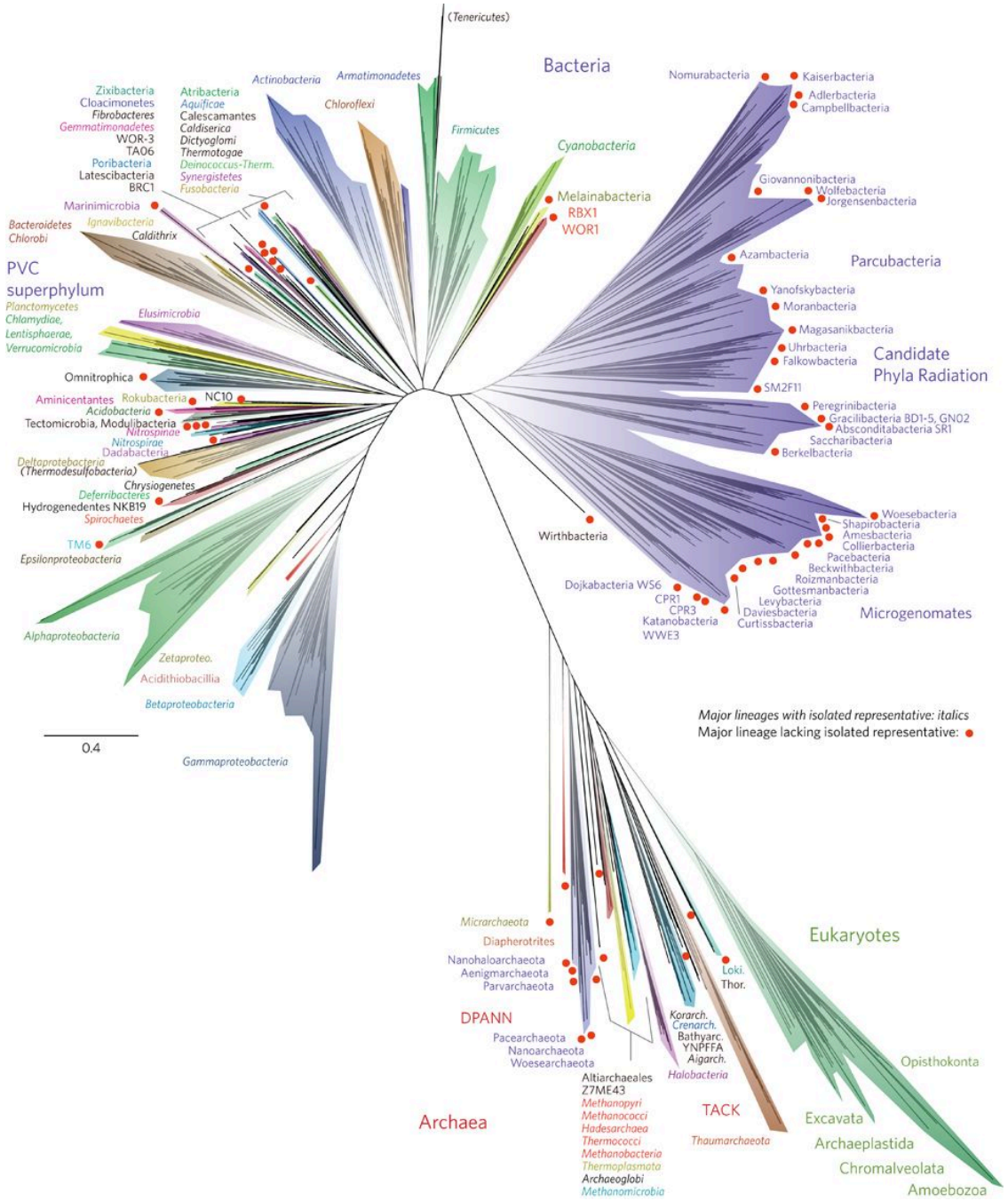
Robert M Bowers, Nikos C Kyrpides, [...] Tanja Woyke

Phylogenetic Classification



Two main competing camps of practice:

- Banfield Lab: manual classification using ribosomal proteins
- Hugenholtz/Ecogenomics Group:



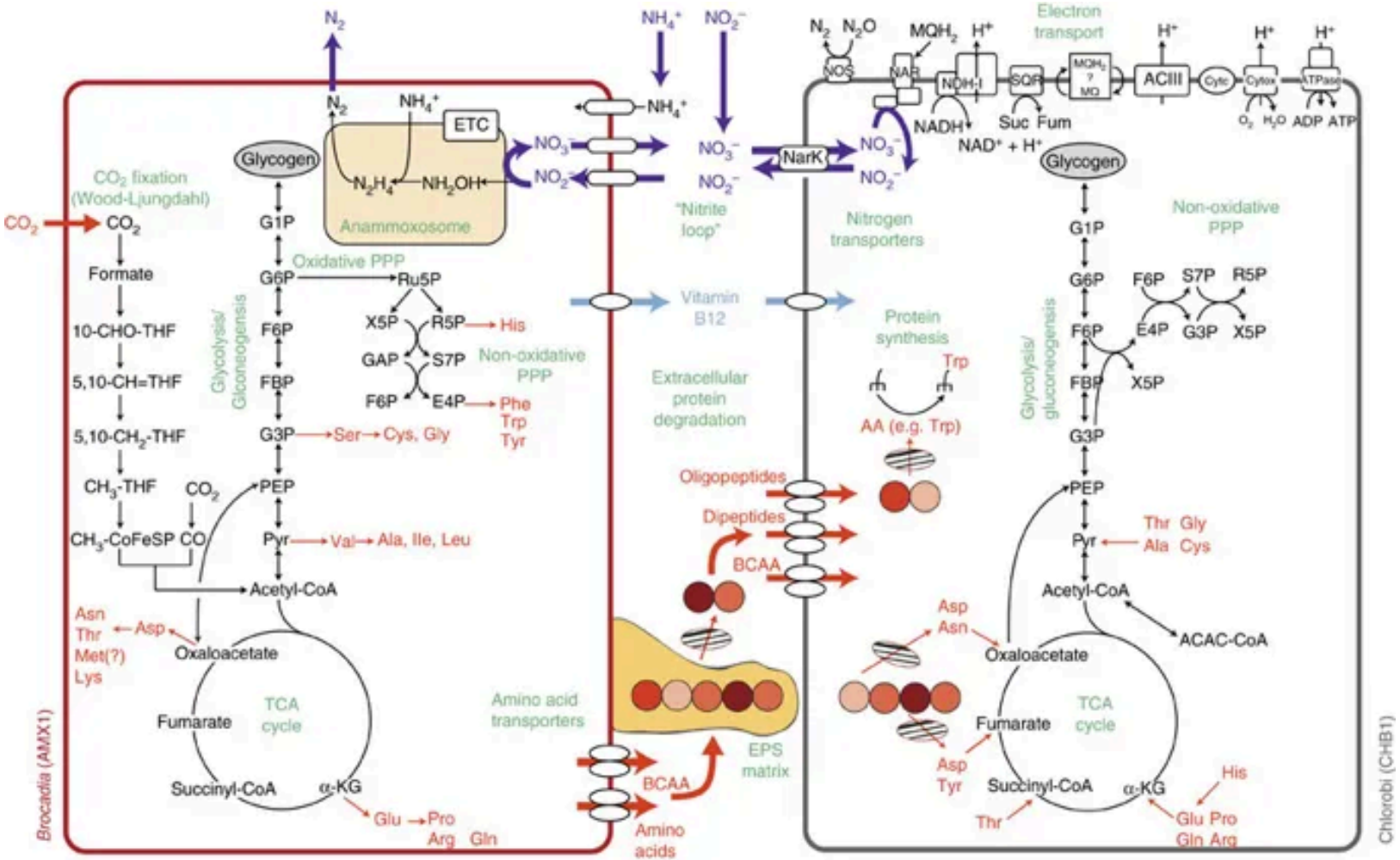
Functional Annotations & Metabolic Reconstructions

Numerous annotation pipelines & databases depending on what your question is

- Prokka: general use
- kofamKOALA: annotate w/ KEGG db
- dbcan: CAZymes
- antiSMASH: secondary metabolites
- MetaPathways: multiple dbs for mapping onto MetaCyc pathways

Mostly all involve a lot of manual parsing

Lawson et al. 2017



Incorporating Long Reads

PACBIO®

Oxford NANOPORE Technologies

A couple of approaches with using long reads:

- Long read assembly only, have to account for high error rates (Canu, Unicycler)
- Hybrid assembly with short and long reads (OPERA-MS)
- Long read assembly polished with short reads @ various stages (Assemble with Canu, polish with Rakon)

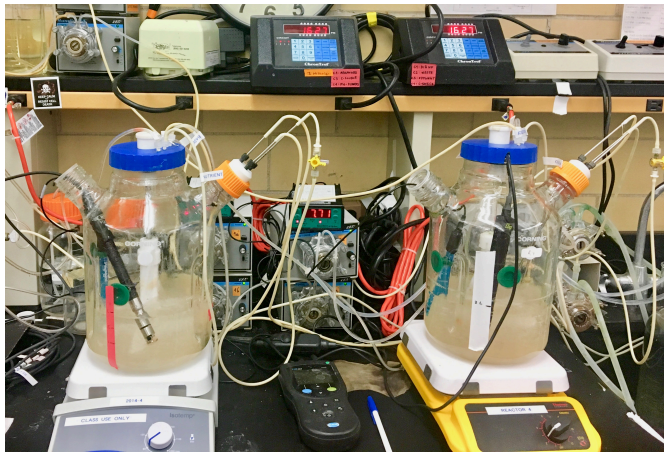
Overall a very new field as long-read technologies get better and error rates go down – current claim of PacBio HiFi reads are 99% accuracy

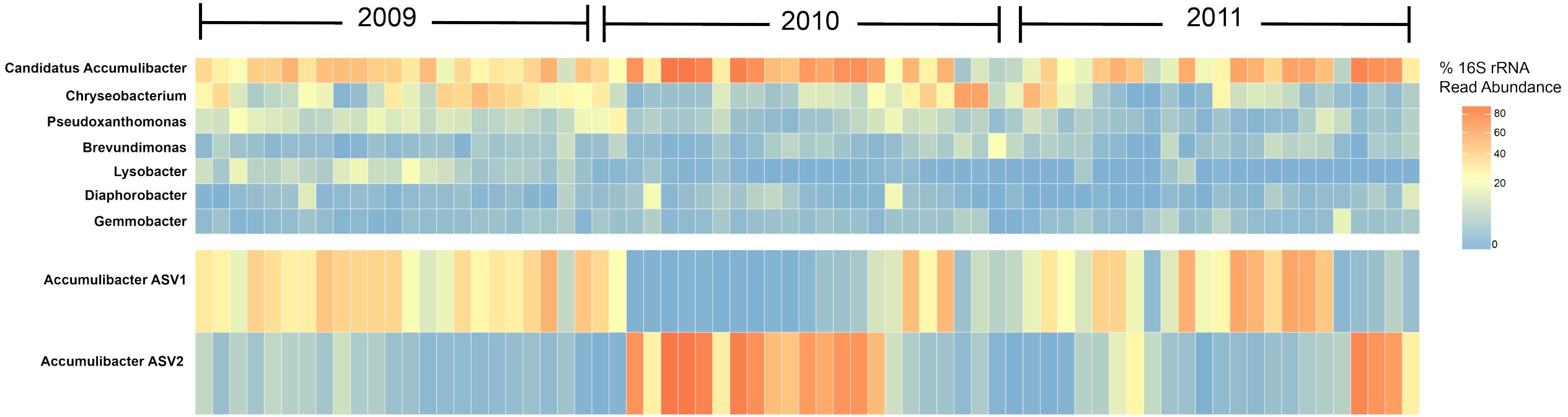
Contamination of Host DNA

Study host-associated microbiomes, oftentimes will get contamination of said host genome that you don't want in your analysis, some approaches for dealing with that:

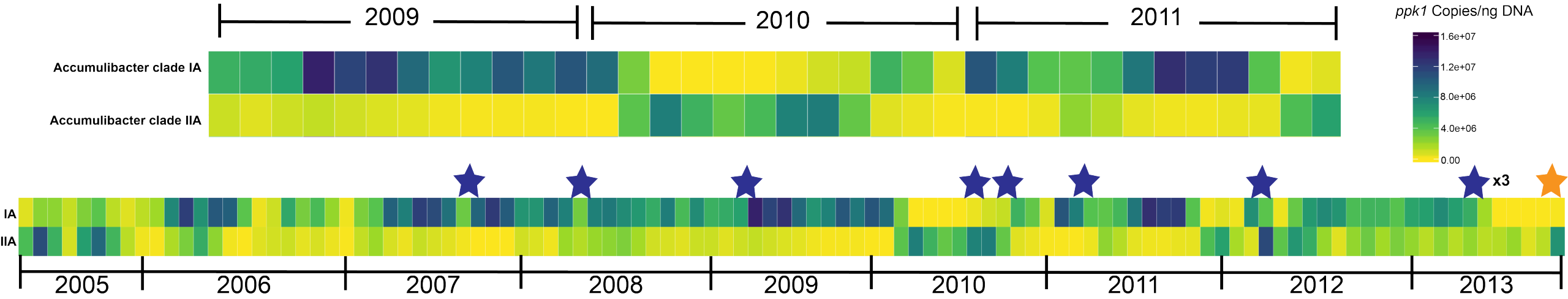
- If have a decent reference genome of your host, map all reads to host, and get rid of those reads (BBMap has a sort specific function for getting rid of contaminating reads based on mapping to a reference genome)
- Anvi'o might also have a workflow for this, depending on whether or not you have a decent reference genome
- AutoMeta metagenomic binning algorithm for dealing with this specific thing, starts with hierarchical taxonomical classification to get rid of host contamination: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkz148

Enrichment Bioreactors Example





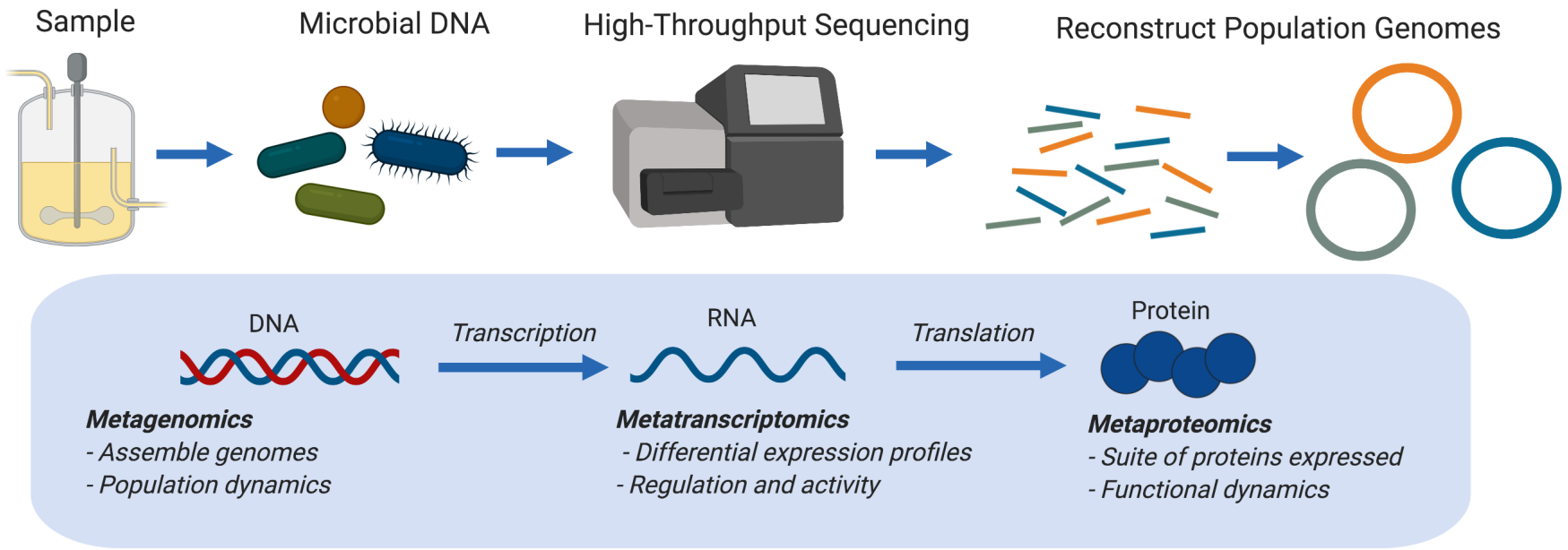
Metagenomic Time-Series



10 metagenomic time-points

6 metatranscriptomic time-points
(Oyserman et al. 2016)

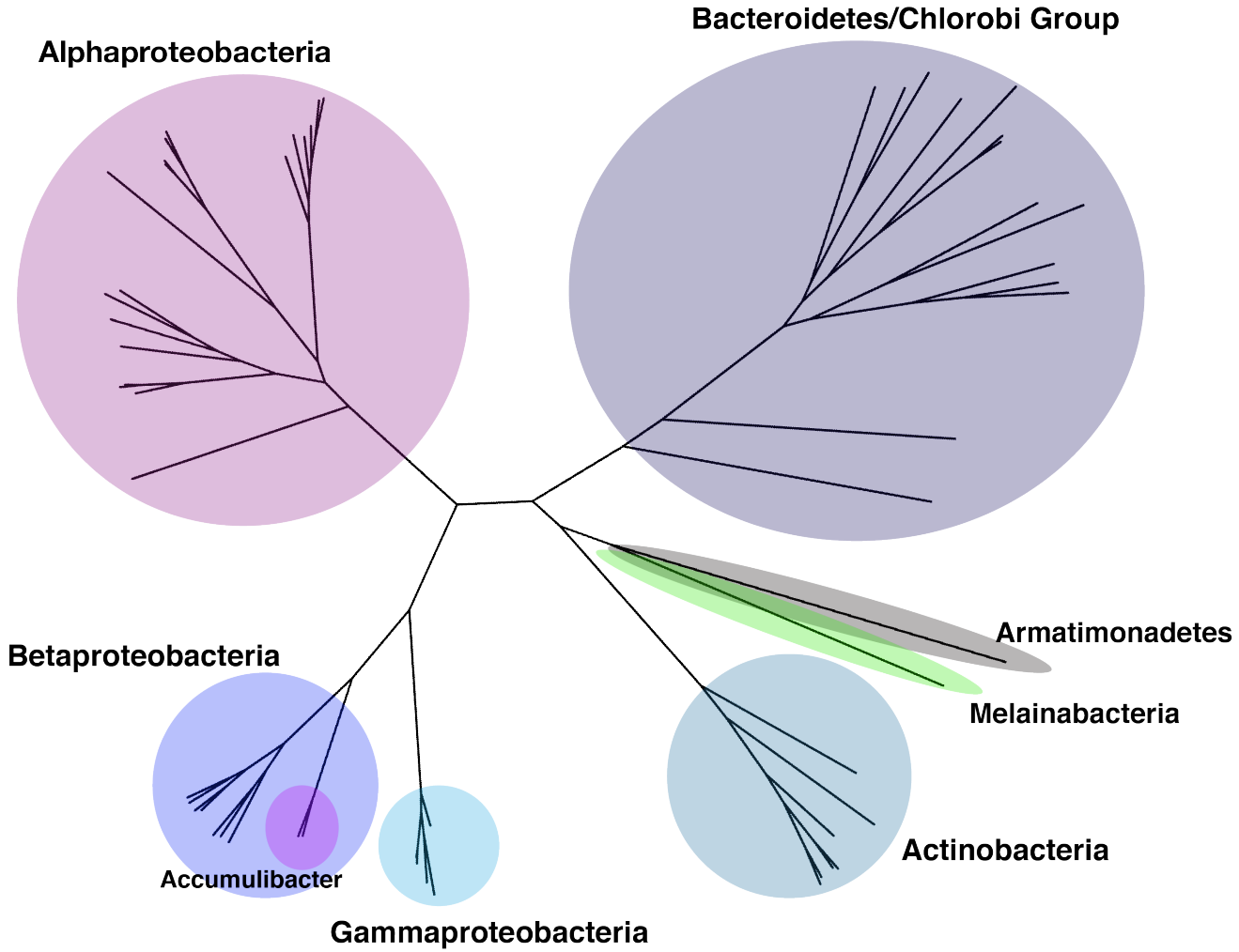
Binning and Dereplication



1. Each metagenomic sample individually assembled with metaSPAdes
2. Each metagenomic time-point mapped reciprocally to each assembly for differential coverage
3. Dereplicated to pick the “best” bin by ANI clusters to pick the most high-quality representative species-resolved MAG
4. Quality check with CheckM, ribosomal markers with HMMs, Barrnap, Infernal
5. Manually inspect uniform coverage using Anvi’o

https://github.com/elizabethmcd/EBPR-MAGs

57 High-Quality* Species-Resolved Genomes



✓ *Overall Stats:*
> 95% complete
< 5% redundant

✓ *tRNAs:*
> 18 tRNA genes

✓ *Individual ribosomal subunits:*
42/57 5S
32/57 16S
25/57 23S

✓ *All 3 ribosomal subunits:*
21/57 with 5S/16S/23S

1. Manually verified classification using ribosomal proteins and reference genomes
2. Automatic taxonomic assignments with GTDB-tk (memory intensive)

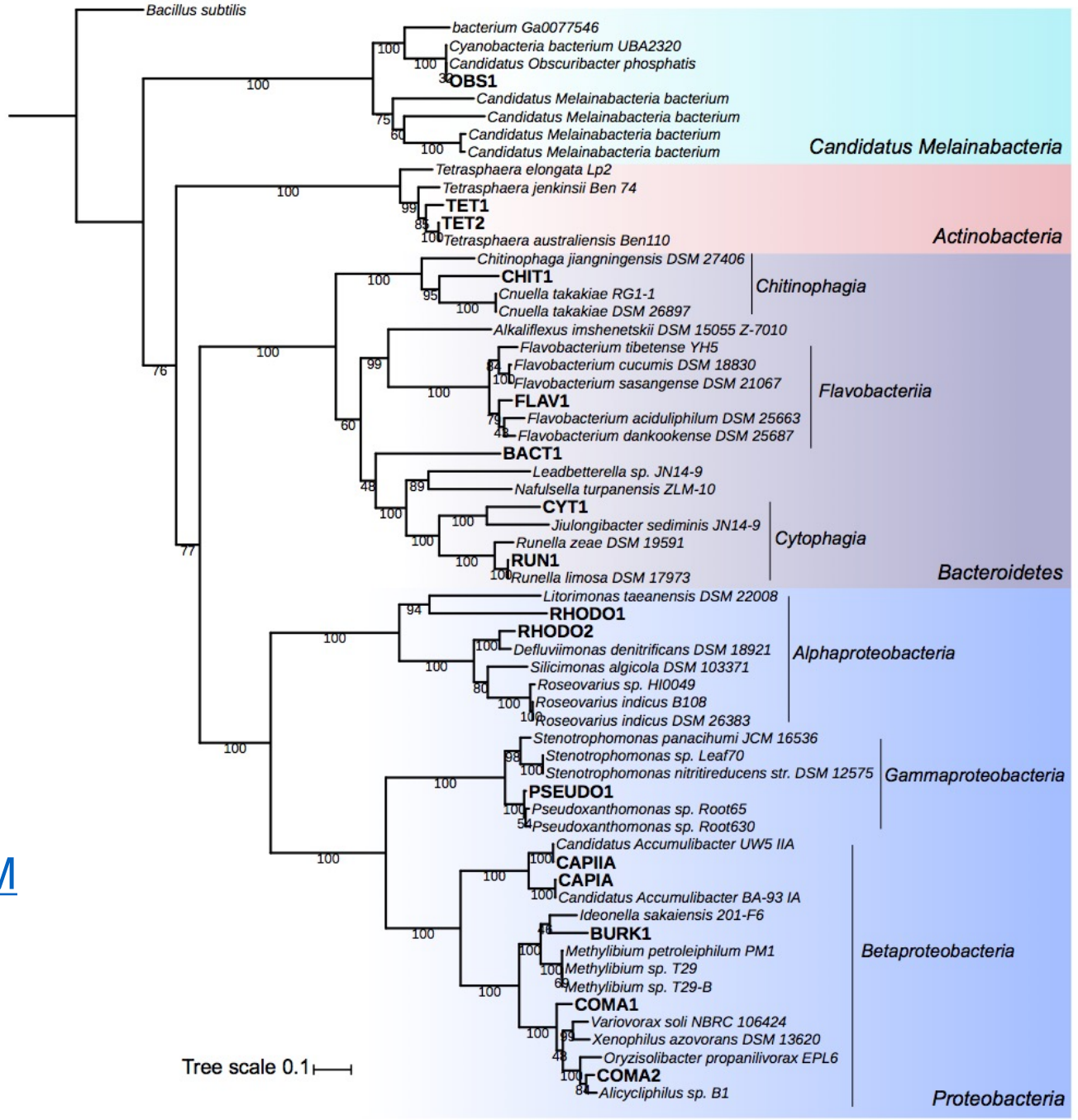
https://github.com/elizabethmcd/EBPR-MAGs

Verifying Taxonomy

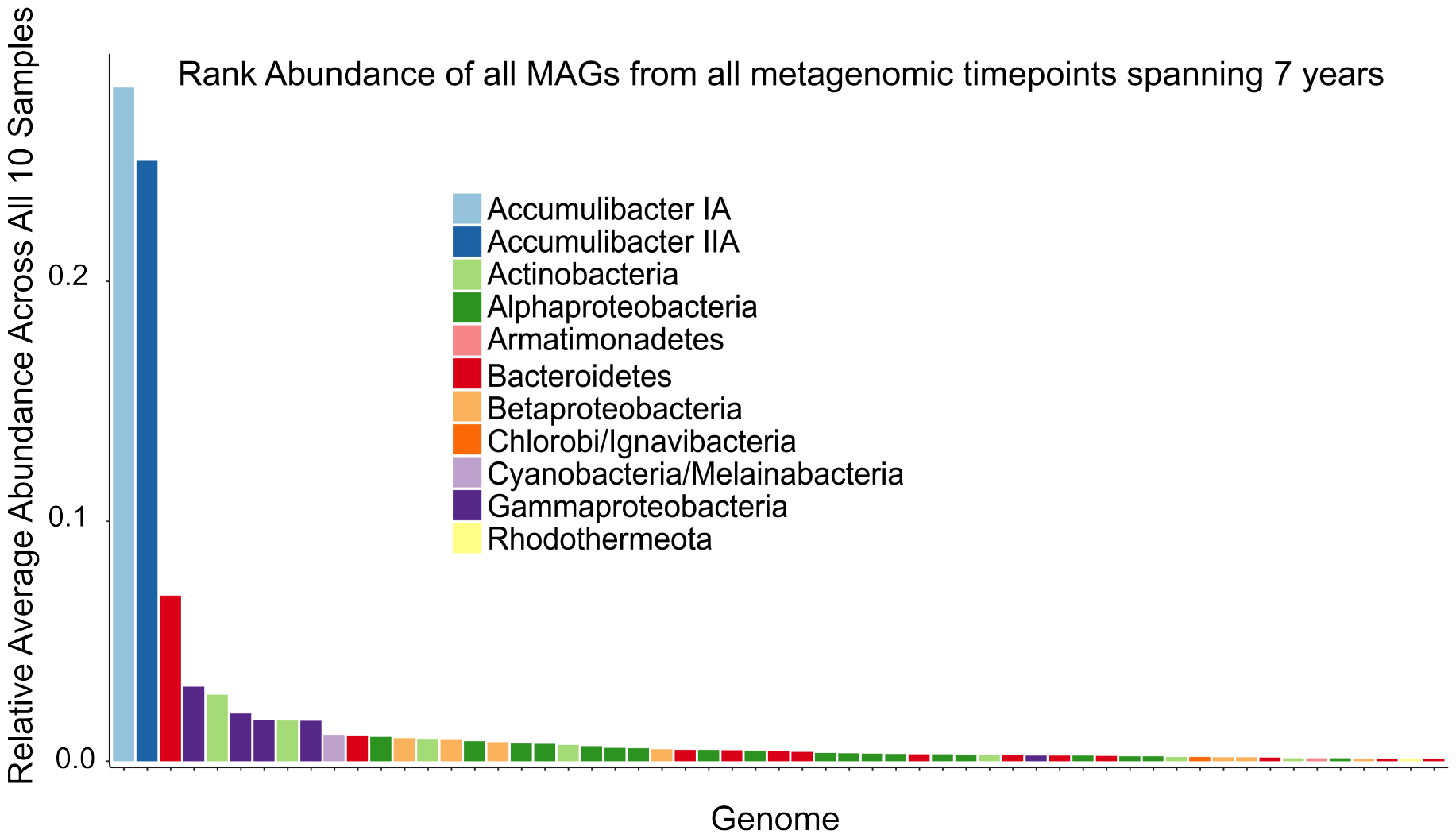
- Either know of existing references you want to compare your new MAGs to, environment-specific databases (MiDAS to activate sludge and anaerobic digestors specific taxonomy db and collection of genomes)

- JSpeciesWS: http://jspecies.ribohost.com/jspeciesws/ using the Tetra Corrleation Search to get rough ballpark estimate if you have no clue

- Make tree of ribosomal proteins, metabolisHMM: https://github.com/elizabethmcd/metabolisHMM



Recovery of Abundant and Rare MAGs across Time-Series



Map metagenomic reads from all samples back to assembled bins to get “relative” abundance, also measure of how well your assembled bins capture sequenced reads

https://github.com/elizabethmcd/EBPR-MAGs

Functional Annotations and Other Considerations

Functional annotations and files in various formats for other downstream steps: **Prokka:** https://www.ncbi.nlm.nih.gov/pubmed/24642063

Biogeochemical summaries based on curated metabolic markers: **METABOLIC:** https://github.com/AnantharamanLab/METABOLIC

One-stop-shop for MAG exploration from taxonomical classifications to functional exploration: **MetaSanity**, with some features in previous **KEGGdecoder:** https://www.biorxiv.org/content/10.1101/789024v1

Genome-resolved metatranscriptomics: **kallisto:** https://pachterlab.github.io/kallisto/

- Combines mapping through pseudoalignment and gene count quantification (previous best practices usually an alignment program of your choice, such as BBMap, and then quantify with HTSeq or featurecounts, all of which are much slower and a pain)
- Easily import into R for downstream exploration and differential expression analyses using the **tximport** package

I attempt to keep reproducible workflows and notebooks of binning enrichment bioreactor and freshwater lake communities on my github: https://github.com/elizabethmcd